# Leveraging Unstructured Information Using Topic Modelling

J.W. Uys[1], N.D. du Preez[1], E.W. Uys[2]
[1]Department of Industrial Engineering, University of Stellenbosch, Stellenbosch, South Africa
[2]Indutech (Pty) Ltd., Stellenbosch, South Africa

**Abstract--Unstructured information in the form of natural language text is abundant in various kinds of organisations. To increase information sharing, organisational learning, decision-making and productivity, large amounts of unstructured text need to be analysed on a daily basis. Full text searching alone is not sufficient as a first approach to help users understand what a collection of electronic documents is about, since it does not provide the user with an overview of the underlying concepts in the document collection.**

**A *topic model* is a useful mechanism for identifying and characterising various concepts embedded in a document collection allowing the user to navigate the collection in a topic-guided manner. Topics, made up of significant words, provide the user with an overview of the content of the document collection. Each document is represented as a mixture of automatically constructed topics and the user may select documents related to a specific topic of interest and vice versa. Similarities between documents may be found by looking at what documents are assigned to a specific topic enabling the user to find other documents related to a given document.**

**This methodology enables users to digest a larger number of documents, assisting them in spending more of their time in actually reading than finding relevant information.**

## I. INTRODUCTION

Electronic text is a convenient and common way to capture and store a variety of information types such as facts, currencies, dates, persons, places, etc. Unstructured information in the form of natural language text is abundant in various kinds of organisations [7] and expands daily. The effort associated with reading and understanding large collections of unstructured information – or natural language text – remains a challenge in spite of many technological advances in the field of information and communication technology [18]. New tools are required for automatically organising, searching, indexing, and browsing the ever-growing, large collections of electronic documents [4]. Currently, unstructured information is found in physical objects such as books, reports, academic dissertations, magazines, and newspapers, to name but a few. Unstructured information is also found in virtual objects such as text messages on cellular phones, web pages (including wikis, blogs, etc.), word processor and other computer files, e-mails, e-books, instant messaging messages, databases (issue tracking systems, customer relationship management systems, etc.) and many more. The amount of time available for an individual to collect, read, interpret and act upon appropriate natural language text is limited both in corporate and research environments. Three (broad) types of technology have been developed to address the problem of working with large collections of textual data typically found in electronic documents [18]:

- Information retrieval technologies,
- Clustering/classification technologies, and
- Natural Language Processing (NLP), data mining, and visualisation technologies

Information retrieval technologies facilitate the process of searching for documents based on user supplied queries and return ranked sets of matching documents as output. Clustering/classification technologies, on the other hand, are mostly used to organise the individual documents of a document collection (corpus), by grouping such documents into representative clusters based on the content of such documents. Some clustering/classification technologies further provide a characterisation of the topics contained in the document corpus. This is achieved by supplementing each of the calculated clusters with a description, consisting of the key terms typifying the relevant cluster. Such descriptions provide a useful overview of the document corpus in terms of the topics addressed as well as the relative coverage of the topics in the corpus. In general, natural language processing (NLP) and data mining technologies have the aim to discover knowledge by extracting, uncovering and synthesising interesting information from document collections.

It is extremely difficult to provide a complete characterisation of well-formed linguistic utterances, since humans bend the rules to satisfy their communicative and creativity needs [14]. This complicates the task of automating understanding and interpreting linguistic textual information. Natural language text is classified as unstructured information due to the absence of explicit structure, found in relational databases and XML, for example. This paper discusses how topic models, essentially a form of clustering/classification technology, could be applied to assist knowledge workers in digesting large collections of textual documents. In the next section, some traditional approaches for analysing large textual document collections are presented. Section 0 lists the shortcomings of the traditional approaches. Section **Error! Reference source not found.** addresses the application of topic models for analysing and characterising document collections. Section 0 provides an example of typical topic model, and discusses the interpretation of such results. The validation of the results of a topic model is briefly discussed in section 0. Section 0 presents ideas around future research, while section 0 gives some concluding remarks.

## II. TRADITIONAL APPROACHES FOR ANALYSING TEXTUAL DOCUMENT COLLECTIONS

Once a set of electronic (textual) documents has been specified by an individual to analyse with a specific objective (e.g. to compile a brief report on the e-health services available internationally), several traditional analysis approaches may be used.

The first approach is to study the abstract and paragraph headings to assess the usefulness of each document. Subsequently, analysed documents can manually be organised into representative categories. Once all documents have been processed in this fashion, the analyst then decides which documents are most relevant and proceeds to evaluate these documents in detail. The entire process relies on user intervention and therefore is time consuming, but accurate.

A second approach is to use a summarisation tool (e.g. *Copernic Summarizer*[1]) or the summarisation functionality of a word processing tool) to reduce the magnitude of the task. Concise summaries of each text are generated, and subsequently scanned by the analyst, to assess the usefulness of each document. Finally, the analyst could classify the documents into categories. Although the second approach is less labour intensive, every document still has to be summarised and evaluated individually. Based on the authors' experience, common text summary tools available today are still largely ineffective in extracting the gist of the natural language text contained in lengthy documents.

A third approach that is widely used to analyse a large set of natural language text documents, is full text indexing the document corpus using software such as *dtSearch*[2], *Lucene*[3], *Google Desktop*[4] . Once indexed, natural language queries similar to that of mainstream web search engines can be used to identify documents matching a specific query. The result set is usually ranked by the software according to the extent in which the individual documents match the supplied query. One can then navigate between the different documents in the result set. Occurrences of the supplied query terms are usually highlighted. The user then repeats this process by entering different queries and assessing the relevant documents returned. This approach generally involves less initial manual effort. It still relies heavily on the appropriateness and quality of the supplied queries. This is especially the case when the analyst is uncertain about what precisely to search for. This limitation is even more severe when the content of the document collection is largely new territory for the searcher [18]. The finding-a-needle-in-a-haystack approach of knowledge retrieval is not always optimal for answering all type of questions [13].

[1] http://www.copernic.com/en/products/summarizer/
[2] http://www.dtsearch.com/
[3] http://lucene.apache.org/java/docs/
[4] http://desktop.google.com/features.html

## III. SHORTCOMINGS OF TRADITIONAL APPROACHES

Traditional approaches for analysing a collection of electronic documents discussed in the previous section have the following limitations:

- There is usually no overview available of the concepts addressed in the corpus or in the individual documents. This makes it difficult to decide which documents to evaluate as well as where to start with the detailed evaluation process.
- The interdependence of the documents constituting the corpus is usually only known after all documents have been analysed by the individual, making it difficult to determine a 'reading path' through such a document collection.
- Documents need to be organised manually using, in most cases, nothing more than human judgement making it a labour intensive process.

Topic models can alleviate these limitations, as explained in section 0.

## IV. INTRODUCTION TO TOPIC MODELS

A *topic model* is a useful mechanism for identifying and characterising various concepts embedded in a document collection allowing the user to navigate the collection in a topic-guided manner [2]. The application of topic models to represent documents has recently received considerable attention in the field of machine learning [21]. Topic models generate interpretable, semantically consistent topics, which can be represented by listing the most probable words describing each topic. A topic model can be defined as a generative model[6] for documents as it specifies a simple probabilistic procedure by which documents can be generated. When compiling a new document, one selects a distribution over topics. Subsequently, for each word in the document, one chooses a topic at random according to this distribution, and selects a word from that topic. Standard statistical techniques may be applied to infer the set of topics that were responsible for generating a collection of documents, thereby reversing the modelled authoring process [20]. Topic models are useful for a variety of tasks such as: organization, classification, collaborative filtering and information retrieval [6]. Topic models generate interpretable, semantically coherent topics, which can be examined by enumerating the most likely words for each topic [16]. Furthermore, topic models are well suited to cater for synonymy (multiple words with similar meanings) and polysemy (words with multiple meanings), since they assign words to topics based on the context of the document [15]. Apart from calculating the topics covered in a collection of documents, topic models further produce a set of individual probabilities that any given document in the collection is about any of the calculated topics. This allows one to learn which documents are significant in terms of which topics. A trained topic model calculates an estimate of the probability

of a word given a topic, $P(w|t)$, and the probability of a topic given a document, $P(t|d)$ for all topics calculated and all documents analysed [15]. Document modelling, corresponding to estimating probability of a word given a document $P(w|d)$, is crucial to information retrieval. Under the LDA model, this probability is not explicitly given, but can be derived from $P(w|t)$ and $P(t|d)$.

## V. OVERVIEW OF TOPIC MODELLING RELATED APPROACHES

Several topic modelling related approaches exists. The following are some of the most popular topic modelling related approaches addressed in information retrieval and machine learning literature [20].

- Latent Semantic Indexing (LSI) [8] [5],
- Mixture of unigrams model, [19], and
- Probabilistic Latent Semantic Indexing (pLSI) [10],
- Latent Dirichlet Allocation (LDA) [1]

Most of these approaches use some kind of dimensionality reduction technique to represent documents using fewer words. LSI uses a linear algebra technique, namely singular value decomposition (SVD) and the bag-of-words representation of text documents for extracting words with similar meanings [9]. LSI is able to model synonymy and polysemy, but has the drawbacks the computation of an SVD does not scale well and that it is a non-probabilistic model. [1] pLSI, the probabilistic variant of LSI, has a statistical foundation and attempts to define a generative data model [6]. A probabilistic model has the advantage that standard statistical techniques can be applied for questions like model fitting, model combination, and complexity control [10]. Although pLSI attempts to relax the simplifying assumption that each document is generated from only one topic, it is not a well-defined generative model of documents, since there is no natural way it can be used to assign probability to a previously unseen document [1]. The Mixture of Unigrams model also employs a probabilistic generative model for the data, but assumes that each document is characterised by exactly one topic [1]. This assumption is regarded to be too simplistic to effectively model a large document collection [21].

LDA, on the other hand, models a document as a mixture of multiple topics and allows documents to exhibit multiple topics to varied degrees and also caters for synonymy and polysemy. Furthermore, LDA is a true generative probabilistic model of a document corpus suitable for the application of statistical techniques. LDA therefore potentially overcomes the drawbacks of earlier topic models (e.g. pLSI, Mixture of Unigrams, etc.) due to its generative properties and the possibility of assigning multiple topics per document. Fundamentally, LDA represents documents as random mixtures over latent topics, where each topic is characterized by a distribution over words in the corpus [1] [15]. The LDA model assumes that the words of each document originate from a mixture of topics, each of which is a distribution over the corpus vocabulary [2]. LDA is a very useful model for deriving structure in otherwise unstructured data as well as generalising new data to fit into that structure [6]. LDA offers a fresh, interesting approach to model documents (compared to the standard query likelihood model) [21].

An even more complex challenge is learning a topic hierarchy from data. Given a collection of documents, each containing a set of words, the challenge in this case is not only to discover the underlying topics in the documents, but further to organize these topics into a hierarchy. An extension of the LDA model, named hierarchical LDA (hLDA), has been developed to do just this [3].

The LDA model assumes that documents are exchangeable[7]. In some cases this assumption is too restrictive, since documents about the same topic are not exchangeable as topics evolve over time. Document collections such as scholarly journals, email, news articles, and search query logs all exhibit evolving content [4]. In another topic modelling approach, named Dynamic Topic Modelling (DTM), the document corpus is divided into sequential segments (e.g., by year) and the document exchangeability assumption is made stricter by assuming that only the documents in each segment are exchangeable [4]. DTM is then applied to the segmented corpus allowing topic distributions to evolve from segment to segment resulting in a hierarchical model of sequential document collections. Dynamic topic models provide a qualitative overview of the contents of a large document collection and also give quantitative, predictive models of a sequential document corpus [4].

Another limitation of LDA is its inability to model correlation between topics. The correlation between topics may be very useful when, for instance, a researcher in a narrow sub-discipline, is searching for a particular research article, and finds that this article is highly correlated with another topic that the researcher may not have been aware about and that is not explicitly included in the article. With the knowledge of the existence of this new related topic, the researcher could explore the document collection in a topic-guided manner to enquire connections to a body of work the researcher was previously unaware of. In most text document collections, it is realistic to anticipate that subsets of the underlying latent topics will be highly correlated. Another topic modelling approach, called the Correlated Topic Model

[5] According to a strict definition of a topic model, LSI is not a true topic modelling technique since it does not represent a generative model for documents.

[6] A generative model is used to randomly generate observed data, usually including some hidden parameters. It defines a joint probability distribution over observation and label sequences. Generative models are applied in machine learning for either directly modelling data, or as an intermediate step to creating a conditional probability density function. A conditional distribution can be created from a generative model through the application of Bayes' rule.

[7] In other words, that joint probability of documents is invariant to permutation.

(CTM), builds on LDA and employs an alternative, more flexible distribution for the topic proportions which provides for covariance structure among topics. This results in a more realistic model of the latent topic structure where the presence of one latent topic may be correlated with the presence of another. CTM provides a natural mechanism for visualising and exploring unstructured data sets [5].

The Pachinko Allocation Model (PAM) [11] presents a flexible alternative to CTM, which has the limitation that it only captures correlations between pairs of topics. PAM captures arbitrary, nested, and possibly sparse correlations between topics using a directed acyclic graph [12]. PAM however, does not represent a nested hierarchy of topics. Another member of the PAM family, hierarchical PAM (hPAM), is an enhanced version of PAM that explicitly represents a topic hierarchy. hPAM combines the advantages of the topical hierarchy representation hLDA with PAM's ability to mix multiple leaves of the topic hierarchy [17].

This concludes the overview of topic modelling related approaches. Non-hierarchical, non-dynamic LDA was used to perform the analysis presented in this paper.

## VI. THE TOPIC MODELLING PROCESS

The high-level topic modelling process, based on the LDA approach, will now be explained.

Firstly, a list of stop words[8] is identified for the corpus to be analysed. All words are then extracted from the document corpus and stop words are eliminated. This results in the corpus vocabulary. The model also maintains a reference of where (i.e. in which document) words occur, and with what frequency. The topic model calculates a number of topics[9], consisting of words as found in the corpus vocabulary. Each word in the corpus vocabulary is then associated with one or more topics with a probability, as calculated by the model. As part of the output of the LDA run, the topic-word matrix presents the topics calculated by the model as well as the words associated with each topic. Each topic is further associated with one or more documents in the collection with a given mixing ratio based on the occurrences of words per document. Also part of the output of the LDA run, the document-topic matrix represents the likely allocations of documents to the topics calculated. On inspection of the resulting topics, each topic may be given a descriptive label by the analyst by evaluating the words and terms allocated to the specific topic (e.g. the label "project management" was given to $topic_j$ in fig. 1).

Fig. 1 below illustrates the concept of a (non-hierarchical, non-dynamic) topic model.

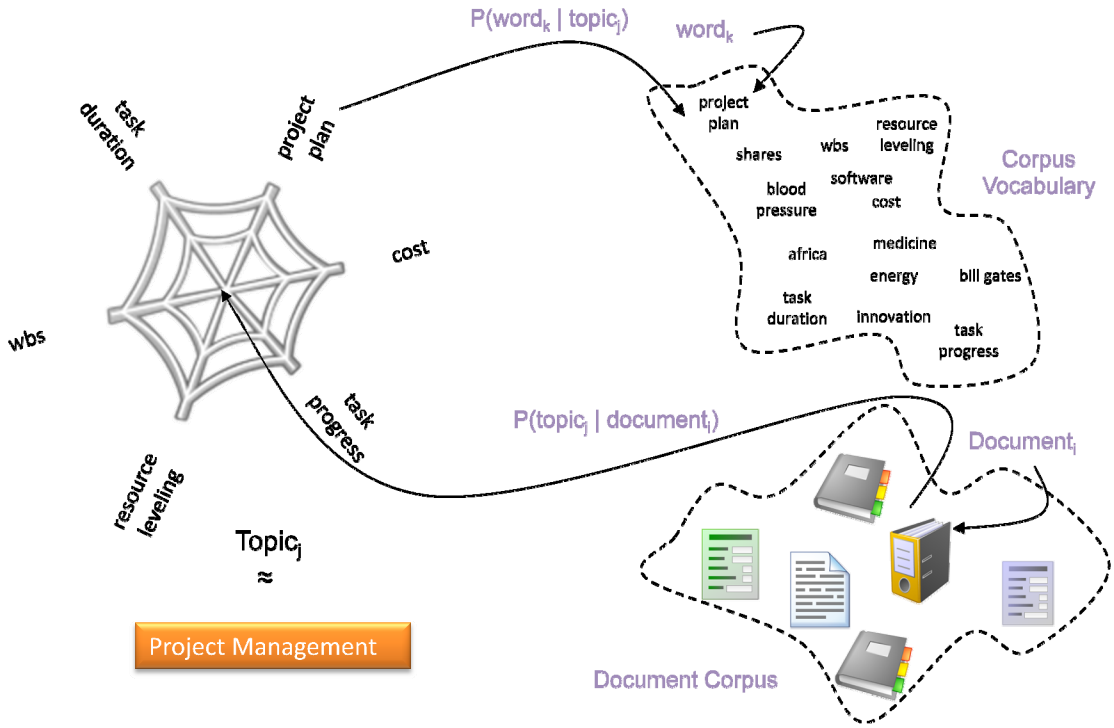


Figure 1: Representation of the Topic Model Concept

[8] Stop words are words having no real significance (e.g. "a", "an", "like", etc.)

[9] The number of topics desired is specified by the user

Since the LDA model usually allocates a given document to more than one topic, the most significant topics associated with a given document can be estimated by considering those topics that corresponds to the given document with the largest mixing ratio. With LDA, the similarity of a given document to other documents can be estimated by identifying other documents that were allocated to the same topic as the given document with high mixing ratios.

## VII. INTERPRETING THE LDA RESULTS: AN EXAMPLE

Table 1 shows the partial topic-word matrix of a 20 topic run on a small document collection (62 documents) addressing a range of health related issues. In-house developed software, implementing the LDA model, was used to do the analysis. Documents were collected from different sources by doing extensive searches using mainstream search engines on the Internet and library repositories. The objective of this topic modelling exercise was to obtain an overview of the types of health information that are addressed by the documents in the relevant corpus, as well as to obtain a list of documents pertaining to the field of mental health. By examining seven out of the twenty topics presented in table 1 below, it is apparent that the analysed documents cover a wide range of health related topics. Each topic in this example is described by 25 words (unigrams or bigrams more specifically) taken from the set of words occurring in the documents corpus. These topics are automatically calculated by the topic model. The most significant word for a given topic is listed first, whereas the less significant words are listed lower down in the relevant columns.

The words in bold were marked by manual inspection of the results and indicate terms that were central in labelling the topic in question. On closer examination of the topics presented, topic 1 apparently deals with mental health disorders, prevention and intervention, whereas topic 2 addresses the treatment and control of malaria. Topic 4 pertains to heart diseases and related factors, while topic 5 deals with obesity. Topic 8 addresses innovation in the medical sector, topic 10 also deals with the treatment of malaria and lastly, topic 14 covers infrastructural issues of urban life.

Although table 1 only presents a partial topic-word matrix, it illustrates that the calculated topics still give a good overview of the concepts underlying the document corpus in question. Note that topics are generally not mutually exclusive with respect to the words that constitute them. However, two topics can never share all of the same descriptive words.

TABLE 1: EXAMPLE OF A TOPIC-WORD MATRIX

| **Topic 1** | **Topic 2** | **Topic 4** | **Topic 5** | **Topic 8** | **Topic 10** | **Topic 14** | **…** |
|---|---|---|---|---|---|---|---|
| **health** | **malaria** | **hbalc key** | **preventie** | **innovation** | **health** | **urban** | … |
| **mental** | **university** | **blood** | **obesity** | **diffusion** | **treatment** | **cities** | … |
| asp xhtml | **study** | **health** | **fatty** | **medical** | **malaria** | **city** | … |
| **prevention** | **health** | **age** | **acids** | **adoption** | **disease** | **water** | … |
| **evidence** | **patients** | **heart disease** | **health** | exchange xml | **medicine** | **regular technology** | … |
| **social** | **africa** | modifications hairpins | **bmi** | diffusion deployment | **patients** | **decent toilet** | … |
| **promotion** | **plasmodium** | **pressure** | wetlands lakes | change | **infection** | **energy** | … |
| **national institute** | **human** | **disease** | **disease** | **technology** | **tropical** | **local** | … |
| **disorders** | **treatment** | **age** | **omega** | **innovations** | **france** | **world** | … |
| **interventions** | september | **heart** | mental | coiwersation web | institute | **people** | … |
| **programmes** | **parasite** | **risk** | **weight** | **process** | research | percent | … |
| cost portable | **infected** | sibling | doctor users | **information** | condition developn | note | … |
| risk | **control** | health | weight waist | **device** | built java | cit | … |
| built java | **france** | messages divided | january riagg | **research** | **falciparum** | state | … |
| miten toimit | **infection** | **diabetes** | assessment | **model** | national | **health** | … |
| research | **Drug** | populations identified | management | **invention diffusion** | areas | treat cure | … |
| effective | pop state | nelson olof | authentication server | **adopters** | group | united states | … |
| technology evaluation | **disease** | father | **diabetes asthma** | **invention** | drug | development | … |
| doctor users | **tropics** | family | living proof | **health** | clinical | areas | … |
| factors | paris france | **sodium** | **waist circumference** | **technologies** | children | **food** | … |
| policy | children | **helminthes laboratory** | naampresentati e january | **deployment** | cases | united | … |
| school | Areas | **laboratory helmminthology** | **waist** | solving strategy | study | washington dc | … |
| proven effective | cases | serving | **circumference** | knowledge | **parasite** | **sanitation** | … |
| community | observed | cme lectures | **overweight** | social | **strains** | tuki selvitys | … |
| .. | … | … | … | … | … | … | … |

The degree of overlap between topics can be calculated in several ways [20]. This, however, falls outside the scope of this paper. Topics 2 and 10, for example, both describe malaria related concepts, and seem to be closely related.

Table 2 (below) shows an extract of the document-topic matrix, as calculated by the topic model, and lists the probability that a specific topic is described by a given document. Probabilities in bold print designate documents that significantly describe a given topic (P > 0.7 was used in this example).

TABLE 2: EXAMPLE OF A DOCUMENT-TOPIC MATRIX

| Documents Analysed | Topic 1 | Topic 2 | Topic 4 | Topic 5 | Topic 8 | Topic 10 | Topic 14 | … |
|---|---|---|---|---|---|---|---|---|
| **Medical device innovation.pdf** | 0 | 0 | 0 | 0 | **0.998** | 0 | 0 | … |
| **Medication Therapy Management Services.pdf** | 0.001 | 0 | 0.003 | 0.03 | 0.005 | 0 | 0.209 | … |
| **Our urban future.pdf** | 0 | 0 | 0 | 0 | 0.001 | 0 | **0.742** | … |
| **Nanoscience and Nanotechnology Research at NIH.pdf** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … |
| **e-health studie TI.pdf** | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | … |
| **Building Collaboration for Clinical Research Networks.pdf** | 0.009 | 0 | 0 | 0 | 0.002 | 0 | 0 | … |
| **Evidence based e-health solution.pdf** | 0.009 | 0 | 0 | 0.002 | 0.002 | 0 | 0 | … |
| **2004 WHO promoting_mental health.pdf** | **0.856** | 0 | 0 | 0 | 0.002 | 0 | 0 | … |
| **Roadmaps for Clinical Practice.pdf** | 0 | 0 | 0 | **0.999** | 0 | 0 | 0 | … |
| **e-health_Rob_Gerrits_kleur_2006.pdf** | 0.012 | 0.002 | 0.002 | **0.924** | 0.001 | 0.002 | 0.002 | … |
| **Roadmap for german health research.pdf** | 0.002 | 0 | 0 | 0 | 0.009 | 0 | 0 | … |
| **Brochure KJL_MD.pdf** | 0.001 | 0 | 0 | 0 | 0 | 0 | 0 | … |
| **2004 WHO Prevention_of_Mental_Disorders.pdf** | **0.998** | 0 | 0 | 0 | 0 | 0 | 0 | … |
| **A Roadmap to the virtual psysiological human.pdf** | 0 | 0 | 0 | 0 | 0.001 | 0 | 0 | … |
| **Controlled_Vocabulary_Bibliography.pdf** | 0.001 | 0 | 0 | 0 | 0.004 | 0 | 0 | … |
| **MedTrop_23nov.pdf** | 0 | 0.327 | 0 | 0 | 0 | 0.241 | 0 | … |
| **Health risk after age 40.pdf** | 0 | 0 | **1** | 0 | 0 | 0 | 0 | … |
| | **6.47%** | **1.26%** | **3.52%** | **6.69%** | **3.61%** | **0.98%** | **3.35%** | … |

The sum of all mixture ratios (MRs) in each row of the document-topic matrix amounts to one, i.e. these are normalised. The normalised sum of the mixture ratios of each column gives an indication of how well each topic is represented in the document corpus. From the topics shown, topic 5 is best covered (6.69%) whereas topic 10 the least covered (0.98%) in the constituent documents. It can further be seen that the document *Health risk after age 40.pdf* is solely allocated to topic 4 (MR=1), while topic 1 is best represented in documents *2004 WHO promoting_mental health.pdf* (MR=0.856) and *2004 WHO Prevention_of_Mental_Disorders.pdf* (MR=0.998). No documents significantly describe topics 2 and 10. Lastly, the fact that documents *Roadmaps for Clinical Practice.pdf* and *e-health_Rob_Gerrits_kleur_2006.pdf* are associated with topic 5 with very high mixture ratios (MR=0.999 and 0.924 respectively) indicates that it is very likely that these two documents are very similar in the concepts they address.

The document-topic matrix in conjunction with the topic-word matrix provides a useful mechanism to do the high-level exploration of a document corpus. The topic-word matrix can be analysed first to get a feel for the range of topics the document corpus contains, without necessarily reading any documents. Next, one can assess which topics are well represented in the document corpus and which are not, by looking at the normalised sum of mixing rations of each topic. A topic of interest can be selected from the topic-word matrix and documents significantly describing this topic can be identified by inspecting the document-topic matrix. A prioritised list of documents on a certain topic can thus easily be obtained, and the process may then be repeated for other topics of interest. The topic model also organises the document collection by associating documents and topics – a process that can take a substantial amount of time when using the manual way.

Lastly, more detailed results can be obtained by doing a run on the subset of documents corresponding to the topics of interest. It is further also possible to analyse a single document using a topic modelling approach given that the document contains sufficient text and a small number of topics are specified in the initial configuration of the run. In this case the topic-word matrix will highlight the topics contained in the document in question, while the document-topic matrix will show how well the document covers the topics calculated.

## VIII. VALIDATING THE RESULTS

Validating the topic model results is not a clear-cut process. In the experience of the authors, the best place to start is to look at the topic-word matrix to determine if the relevant terms defining the topics more or less describe an apparent, unambiguous theme for most topics. If this is not the case, the parameters of the model may be adjusted (e.g. use 10 topics instead of 20) and the run repeated. The real test for the results is to scan through a sample of the significant documents per topic to verify that it actually pertains to the topic in question. More formal validation techniques will be investigated in future.

## IX. FUTURE WORK

Further research will include conducting an experiment to analyse several documents of a number of authors, with the goal to calculate a topic profile per author. Furthermore, finding a heuristic to calculate the optimal number of topics for a given document collection would be very useful, or, alternatively, using a non-parametric model, as mentioned in [12]. Currently, the user has to specify the number of topics desired in the output of the topic model. Experimentation with the even more advanced topic modelling techniques described earlier (e.g. hLDA, DTM, CTM, PAM and hPAM) are also planned. Furthermore, visualising the topic model results in a manner that would allow the user to explore a document corpus in an interactive, topic-guided fashion would further receive attention. The research presented here forms part of a PhD study having the aim of developing a dynamic framework to connect the various information entities of an organisation to support the process of intra-organisational innovation.

## X. CONCLUSION

Analysing an electronic document corpus to get an overview of the concepts addressed in the corpus, as well as in the individual documents, can be a tedious process when done by means of traditional approaches such as manual evaluation. This paper illustrated - by means of a simple example - how a topic modelling approach such as LDA may aid an individual to achieve this objective. It was shown how topic models can give an overview of the topics underlying a document collection, how topics can be used to select documents of interest for further processing, and how such a model can be used to detect similarities between documents.

## REFERENCES


[1] Blei, D., Ng, A. and Jordan, M. "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993 – 1022, 2003.

[2] Blei, D. and Lafferty, J. "A Correlated Topic Model of Science," *The Annals of Applied Statistics 2007*, vol. 1, no. 1, pp. 17–35, 2007.

[3] Blei, D., Griffiths, T., Jordan, M. and Tenenbaum, J. "Hierarchical Topic Models and the Nested Chinese Restaurant Process," *NIPS*, 2004.

[4] Blei, D. and Lafferty, J. "Dynamic Topic Models," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[5] Blei, D. and Lafferty, J. "Correlated topic models," in *Advances in neural information processing systems*, vol. 18, 2006.

[6] Blei, D., and Lafferty, J. "Modeling Science," 2006.

[7] Cheung, C., Lee, W. and Wang, Y. "A multi-facet taxonomy system with applications in unstructured knowledge management," *Journal of Knowledge Management*, vol. 9, issue 5, pp. 76 – 91. 2005.

[8] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, issue 6, pp. 391 – 407, 1990.

[9] Fortuna, B., Grobelnik, M. and Mladenić, D. "Visualization of text document corpus," *Informatica,* vol. 29, pp. 497–502, 2005.

[10] Hofmann, T. "Probabilistic latent semantic indexing," *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA, USA. 1999.

[11] Li, W. and McCallum, A. "Pachinko allocation: DAG-structured mixture models of topic correlations," *ICML*, 2006.

[12] Li, W., Blei, D. and McCallum, A. "Nonparametric Bayes pachinko allocation," in *The 23rd Conference on Uncertainty in Artificial Intelligence*, 2007.

[13] Lieberman, J. "From Metadata to Megadata: Deriving Broad Knowledge from Document Collections," Collexis, Inc., Whitepaper. 2007.

[14] Manning, C. and Schütze, H. "Foundations of Statistical Natural Language Processing," MIT Press, Cambridge, MA. 1999.

[15] Mimno, D. and McCallum, A. **"**Expertise Modeling for Matching Papers with Reviewers," *Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '07), San Jose, California, USA, August 2007.

[16] Mimno, D. and McCallum, A. **"**Mining a Digital Library for Influential Authors," *Joint Conference on Digital Libraries* (JCDL '07), Vancouver, British Columbia, Canada, June 2007.

[17] Mimno, D., Li, W.and McCallum, A. "Mixtures of Hierarchical Topics with Pachinko Allocation," in *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[18] Nasukawa, T. and Nagano, T. "Text Analysis and Knowledge Mining System," *IBM Systems Journal*, vol. 40, no. 4, 2001.

[19] Nigam, K., McCallum, A., Thrun, S. and Mitchell, T. "Text classification from labeled and unlabeled documents using EM," Machine Learning, vol. 39, no. 2/3, pp. 103–134, 2000.

[20] Steyvers, M. and Griffiths, T. "Probabalistic Topic Models," *in Latent Semantic Analysis: A Road to Meaning, Trends in Cognitive Science. v*ol. 10, issue 7, pp. 327 – 334, July 2006.

[21] Wei, W. and Croft, W. **"**LDA-Based Document Models for Ad-hoc Retrieval," in *Proceeding of 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval* (SIGIR '06), Seattle, WA, USA, August 2006.